Algoritma K-NN untuk klasifikasi dataset Covid-19 survillance

Ivandari
STMIK Widya Pratama Pekalongan
Email: ivandarialkaromi@gmail.com

RINGKASAN

Covid-19 merupakan jenis virus mutasi baru yang banyak ditemukan dan diteliti di seluruh dunia. Untuk sementara belum ditemukan obat yang efektif untuk mengobati atau mencegah penyakit tersebut. Salah satu cara yang dilakukan berbagai pemerintahan di dunia adalah membatasi kontak fisik dengan penderita covid-19. Data mining adalah salah satu ilmu computer untuk mempelajari data dan menlakukan ekstraksi untuk mendapatkan sebuah pengetahuan baru. Salah satu teknik dalam data mining adalah klasifikasi. K-NN adalah salah satu algoritma klasifikasi terbaik. Penelitian ini melakukan klasifikasi dataset Covid-19 *survillance* menggunakan algoritma K-NN. Dataset Covid-19 *survillance* didapatkan dari portal data public yaitu *uci machine learning repository*. Hasil klasifikasi dengan menggunakan aplikasi bantu rapid miner menghasilkan tingkat akurasi dari K-NN adalah 55%. Tingkat akurasi 55% tergolong dalam tingkat akurasi yang rendah. Rendahnya tingkat akurasi ini dapat disebabkan oleh sedikitnya atribut yang digunakan dalam klasifikasi K-NN, serta adanya dominasi dari salah satu varian dalam atribut label.

Kata Kunci : KNN, covid-19 *survillance*, akurasi

1. PENDAHULUAN

1.1 Latar Belakang

Coronavirus Diseases atau sering lebih dikenal dengan covid-19 merupakan penyakit yang sangat mudah menular dengan penyebaran yang luas dan menjadi isu global saat ini. Di Indonesia kasus Covid-19 pertama kali ditemukan pada awal tahun 2020. Dengan penyebaran yang tergolong sangat cepat dalam bulan pertama kasus terkonfirmasi positif Covid-19 mencapai 1790 orang. Selanjutnya pada akhir 2020 kasus ini meluas sampai ke daerah dengan total kasus mencapai setengah juta jiwa terkonfirmasi positif.

Kebijakan pemerintah dengan membatasi kerumunan serta memberikan vaksinasi terhadap pekerja public dan lansia menjadi salah satu usaha untuk penekanan peningkatan kasus Covid-19. Data mining merupakan satu bidang ilmu untuk mendalami data dan melakukan perhitungan untuk mendapatkan pengetahuan baru dari data tersebut (Witten et al., 2011). Klasifikasi merupakan bagian terpenting dari data mining. Beberapa algoritma klasifikasi digunakan untuk memecahkan suatu maslah. KNN (K-Nearest Neighbour) merupakan salah satu algoritma klasifikasi terbaik (Alkaromi, 2014), dan mudah digunakan (Wu, 2009).

Uci Machine Learning Repository merupakan sekumpulan basis data yang sering digunakan untuk melakukan uji coba sebuah metode atau algoritma. Data dari *Uci Machine Learning Repository* ini merupakan data public yang teruji di segala bidang. Untuk bidang kesehatan sendiri terdapat beberapa antara lain data penderita diabetes, kanker payudara, serta kasus Covid-19 yang merupakan tren penelitian di beberapa tahun ini. Dataset Covid-19 yang diambil di website uci ini merupakan data penderita Covid-19 dengan berbagai gejalanya yang diambil dari Kementrian Kesehatan Republik Indonesia (https://archive.ics.uci.edu/ml/datasets/COVID-19+Surveillance).

Penelitian ini menggunakan algoritma KNN untuk melakukan klasifikasi datset covid-19. Perhitungan dilakukan dengan menggunakan aplikasi bantu yaitu rapid miner. Proses validasi perhitungan dilakukan dengan menggunakan *10 folds cross validation*. Model validasi ini merupakan validasi yang paling banyak

digunakan untuk perhitungan klasifikasi. Perhitungan akurasi algoritma dilakukan dengan menggunakan matrix kebingungan (*confussion matrix*). Dalam proses ini akan diketahui banyaknya *record* yang untuk data uji yang sesuai dengan hasil aslinya. Selanjutnya dilakukan pembagian dengan keseluruhan proses perhitungan sehingga didapatkan nilai prosentase akurasi dari sebuah algoritma.

2. METODE PENELITIAN

Dalam penelitian ini metode eksperimental digunakan untuk memperoleh hasil terbaik klasifikasi dataset Covid-19 menggunakan algoritma KNN. Eksperimen dilakukan untuk memperoleh nilai k terbaik. Tahapan dalam penelitian yang dilakukan antara lain sebagaimana berikut:

2.1. Pengumpulan Data

Proses pengumpulan data dilakukan dengan menggunakan data public. Dalam data yang diterbitkan oleh *uci machine learning repository* ini diperoleh satu dataset dengan 7 atribut regular dan 1 atribut tujuan atau atribut label. Dataset dengan nama Covid-19 survillance ini dapat dilihat serta diunduh dari url: https://archive.ics.uci.edu/ml/datasets/COVID-19+Surveillance.

2.2. Perhitungan Algoritma

Tahapan ini merupakan tahapan pokok penelitian. Dalam tahapan ini terbagi lagi menjadi beberapa bagian, yaitu tahap validasi dan tahap evaluasi.

2.3. *10 folds cross validation*

Tahapan ini dilakukan dengan menggunakan aplikasi bantu yaitu rapid miner. Proses validasi dilakukan dengan membagi dataset menjadi 10 bagian lalu 1 bagian dijadikan sebagai data uji dengan 9 bagian lainnya sebagai pembanding. Proses ini diulang sampai dengan 10 kali dimana bagian lain berubah menjadi data uji pada perulangan berikutnya. Selanjutnya proses perulangan berakhir apabila semua *record* sudah mendapat bagian satu kali menjadi data uji. Representasi *10 folds cross validation* dapat dilihat sebagaimana gambar 1 berikut.

	100% dataset *record*									
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
perulangan 1										
perulangan 2										
perulangan 3										
perulangan 4										
perulangan 5										
perulangan 6										
perulangan 7										
perulangan 8										
perulangan 9										
perulangan 10										

Gambar 1. Representasi *10 folds cross validation*

2.4. *Confussion matrix*

Matrix kebingungan atau biasa juga disebut dengan *confussion matrix* merupakan salah satu model evaluasi yang banyak digunakan dalam penelitian data mining, utamanya di bidang klasifikasi dan prediksi. Proses ini sebenarnya membandingkan data prediksi atau klasifikasi dengan data asli. Perhitungan sederhananya adalah jumlah data prediksi atau klasifikasi yang sesuai dengan kondisi asli dibagi dengan keseluruhan total perhitungan yang dilakukan. Hasil akhir dari perhitungan akan didapatkan output presentase kebenaran. Gambar 2 merupakan representasi proses perhitungan matrix kebingungan.

Classification		*Predicted class*	
		Class: YES	*Class*: NO
Observed class	*Class*YES	***a*** ***True Positive (TP)***	***b*** ***False Negative (FN)***
	*Class*NO	***c*** ***False Positive (FP)***	***d*** ***True Negative (TN)***

Gambar 2. *Confussion matrix* (Gorunescu, 2011)

3. HASIL DAN PEMBAHASAN

3.1. Analisa Atribut Data

Tahap awal yang dilakukan adalah mendapatkan database Covid-19 survillance yang telah didapatkan sebelumnya dari url:

https://archive.ics.uci.edu/ml/datasets/COVID-19+Surveillance. Dalam dataset terdapat 7 atribut regular dan 1 atribut label atau atribut tujuan yaitu *categories*. Semua nama atribut regular sengaja dikodekan untuk menjaga objektifitas dataset. Terdapat 14 *record* dalam dataset ini. Tabel 1 merupakan dataset yang digunakan dalam penelitian ini.

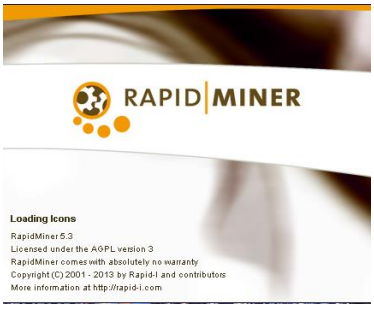
Tabel 1. Dataset Covid-19 survillance

A01	A02	A03	A04	A05	A06	A07	Categories
+	+	+	+	+	-	-	PUS
+	+	-	+	+	-	-	PUS
+	+	+	+	-	+	-	PUS
+	+	-	+	-	+	-	PUS
+	-	-	-	-	-	+	PUS
+	+	+	-	-	-	+	PUS
+	+	-	-	-	-	+	PUS
+	+	+	+	-	-	-	PUS
+	-	-	+	+	-	-	PIM
-	+	-	+	+	-	-	PIM
+	-	-	+	-	+	-	PIM
-	+	-	+	-	+	-	PIM
-	+	-	-	-	-	+	PIM
-	-	-	-	-	-	+	PWS

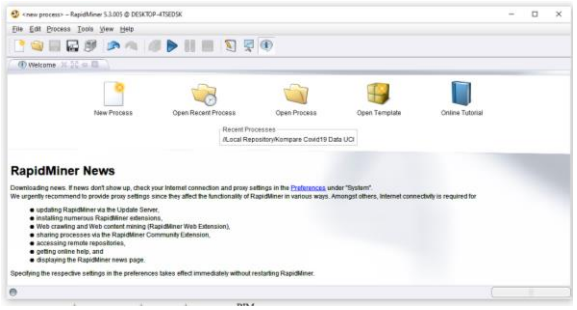
Dari table 1 dapat diketahui bahwa data yang ada menggunakan tipe nominal. Tipe nominal ini adalah semua isian dari *record* data yang tidak dapat di bandingkan menggunakan skala denga isian *record* yang lainnya. Untuk menangani data dengan tipe nominal dapat digunakan berbagai macam metode klasifikasi, termasuk diantaranya adalah K-NN (Alkaromi, n.d.).

3.2. Perhitungan Algoritma

Proses perhitungan dilakukan dengan menggunakan aplikasi bantu yaitu rapid miner. Aplikasi ini banyak digunakan untuk proses pengukuran akurasi sebuah metode data mining. Gambar 3 dan gambar 4 merupakan tampilan utama pada rapid miner.

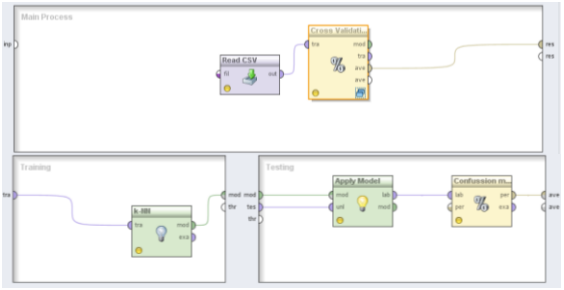


Gambar 3. Tampilan awal rapid miner



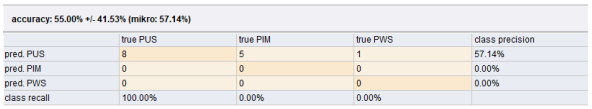
Gambar 4. Halaman utama rapid miner

Dalam proses validasi dilakunan dengan memanfaatkan *10 folds cross validation* serta menggunakan *confussion matrix* untuk melakukan evaluasi dan perhitungan algoritma dengan metode K-NN. Gambar 5 merupakan rangkaian proses dalam aplikasi rapid miner.



Gambar 5. Proses perhitungan

Dari rangkaian proses yang dilakukan sebagaimana pada gambar 5 tersebut, didapatkan hasil *confussion matrix* sebagaimana telihat pada gambar 6.



accuracy: 55.00% +/- 41.53% (mikro: 57.14%)

	true PUS	true PIM	true PWS	class precision
pred. PUS	8	5	1	57.14%
pred. PIM	0	0	0	0.00%
pred. PWS	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

Gambar 6. Hasil perhitungan dari rapid miner

Hasil akurasi dari perhitungan algoritma menggunakan rapid miner menunjukkan tingkat akurasi K-NN adalah 55% dengan rincian sebagaimana pada gambar 6. Dengan akurasi 55% proses klasifikasi ini tergolong dalam tingkat akurasi rendah.

3.3. Pembahasan

Tingkat akurasi yang didapatkan dari proses perhitungan sebelumnya tergolong rendah. Hal ini dikarenakan dataset Covid-19 survillance hanya memiliki 14 *record*. Algoritma K-NN sangat peka terhadap data. Banyaknya *record* dapat mempengaruhi proses pembelajaran pada algoritma K-NN yang hasilnya juga dapat mempengaruhi tingkat akurasi algoritma K-NN (Indrayanti et al., 2017).

Selain dipengaruhi oleh banyaknya *record* yang ada, rendahnya tingkat akurasi juga dapat dipengaruhi oleh rendahnya varian dari atribut label atau atribut tujuan. Dalam dataset Covid-19 survillance terdapat 3 varian label dengan rincian: 8 *record* (PUS), 5 *record* (PIM), serta 1 *record* (PWS). Jumlah ini jelas mempengaruhi hasil klasifikasi karena salah satu varian label sangat dominan dibandingkan varian yang lain.

4. SIMPULAN DAN SARAN

4.1. Simpulan

Hasil dari penelitian ini adalah sebagaimana berikut:

1. Klasifikasi dataset Covid-19 *survillance* menggunakan algoritma K-NN memperoleh tingkat akurasi sebesar 55% yang tergolong dalam tingkat akurasi rendah.
2. Rendahnya tingkat akurasi disebabkan karena sedikitnya *record* yang ada, serta terdapat varian dari atribut label yang dominan.

4.2. Saran

Dalam proses penelitian ini menggunakan dataset yang memiliki 14 *record*, dengan 3 varian atribut label. Salah satu varian atribut label bahkan memiliki jumlah lebih dari 50% dari total *record* yang ada. Penelitian selanjutnya dapat menggunakan dataset dengan jumlah *record* yang lebih banyak.

DAFTAR PUSTAKA

Alkaromi, M. A. (n.d.). *Komparasi Algoritma Klasifikasi untuk dataset iris dengan rapid miner*. *0285*.

Alkaromi, M. A. (2014). *Information Gain untuk Pemilihan Fitur pada Klasifikasi Heregistrasi Calon Mahasiswa dengan Menggunakan K-NN*.

Gorunescu, F. (2011). *Data Mining: Concepts; Models and Techniques*. Springer.

Indrayanti, I., Devi, S., & Al Karomi, M. A. (2017). Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur Gain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus. *IC-TECH*, *XIII*(2), 1–6.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition* (Vol. 40, Issue 6). Elsevier. https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

Wu, X. (2009). *The Top Ten Algorithms in Data Mining* (V. Kumar (ed.)). Taylor & Francis Group, LLC.